# *Thematic vs. social networks in web 2.0 communities: A case study on* Flickr *groups*

Nicolas Pissard[1] and Christophe Prieur[2] †

*Orange France Telecom R&D, 38 av. General Leclerc, 92794 Issy-les-Mx*
*and [1]University Paris Dauphine, [2]Liafa, University Paris Diderot & CNRS*

L'objectif de cet article est d'étudier les groupes d'utilisateurs sur le site web "2.0" *Flickr*, pour déterminer s'ils constituent de réelles "communautés" ou bien des clusters à visée essentiellement thématique. Nous décrivons un cadre méthodologique pour l'analyse de réseaux d'utilisateurs permettant de mesurer si un groupe donné est plutôt thématique, social ou les deux. Des résultats illustratifs sont donnés sur un échantillon de 450 groupes comportant environ 500 membres chacun.

**Keywords:** grands réseaux d'interactions, web 2.0, communautés, réseaux sociaux, folksonomies, flickr

## 1 Introduction

As publishing tools became more accessible and types of online contents more various, the usage of the worldwide web has changed. Rather than just browsing information published by some happy few, today internet users have many occasions to produce information in a much richer way than 'classical' forums. *User-generated content* can take mupltiple forms: sophisticated texts (*Wikipedia*, blogs), photographs (*Flickr*) or video (*Youtube*) [CJPCP06]. As a corollary to this evolution, a general tendency to link internet users to one another has developped (via the 'contacts' functionality on most web services), encouraged by the 6-degrees small-world fantasy [Mil67, Wat03]. If interesting content can come from anyone, you have to build connections to interesting people in order to keep aware of what is coming. The fact that 'interesting' has different meanings for different people yields the notion of 'communities' as a mainstring of this new (so-called 2.0) web.

Identifying these so-called communities or *clusters* as densely connected subgraphs in a network is one of the main issues in complex networks analysis [New04, LP04] already addressed in the times of the good old web 1.β with information retrieval motives [BP98, FLG00], and even before in the first ages of social network analysis [WBB76, WF94] in modeling purposes. What is more difficult to understand is the nature of actual social relations between the members of such a 'community'. Understanding it would help identifying needs for *e.g.* new web applications, infrastructures or business models. Some recent studies address on large data the issue of what is inside a community, especially for blogs [LWGS06, AHA07].

The aim of the present paper is to study user-created groups as communities on the *Flickr* photo publishing website in order to determine whether they are *social media* tools or rather clusters with a mainly thematic purpose. We describe an analysis framework on *Flickr*-users networks that produces a characterization of *Flickr* groups in terms of thematic and/or social aspects, and give results on a sample of 450 groups with around 500 members each. The analysis relies on a graph model using measures inspired from collaborative filtering (see *e.g.* [BHK98]).

†This work is part of *Autograph* (`http://autograph.fing.org`), a project supported by the French ANR/Telecom.
The data was collected by Pascal Pons, with whom we've had many discussions on ways to define metrics between users *via* tags.

# 2 Data

## 2.1 Flickr: photo archive or social media?

*Flickr*[3] is a website that enables users to upload photos, index them with free keywords called *tags* (*e.g. cat*, *paris* etc.) and post them to thematic user-created *groups* (*e.g. Cats rule*, *People in the street* etc.). They can also put *comments* on other users' photos, mark them as their *favorites* and mark these users as their *contacts*.

In addition to its photo pool, each group has a discussion forum which encourages social activity. The great thematic redundancy of these groups (more than 300 groups just about cats) suggests that this social aspect is at least as important as the thematic one[4]. We will thus take into account both thematic (tags) and social (contacts and comments) functionalities used by members of a group to characterize its type. Note that this group feature is what makes *Flickr* a very interesting example for studying existing communities rather than trying to infer them from network structure.

## 2.2 Social and thematic graphs

Let us denote by $U$ the set of all *Flickr* users having at least one photo with at least one tag, by $T$ the set of all tags used on *Flickr* and by $\Gamma$ the set of *Flickr* groups on which we worked (all 450 groups having between 433 and 500 members at the time of the crawl[5]).

Given a group $g \in \Gamma$, we will denote by $U(g)$ the set of all *Flickr* users having posted in $g$ at least one photo with at least one tag. The *social graph* $G_s(g)$ of $g$ is the graph with set of vertices $U(g)$ and set of (undirected) edges $E_s(g)$ such that $u - v \in E_s(g)$ in at least one of the following cases: $u$ is marked as $v$'s contact, $v$ has posted a comment on one of $u$'s photos, or the converse of one of these two[6].

The *thematic graph* $G_t(g)$ is defined with set of vertices $U(g)$ and set of (undirected) edges $E_t(g)$, such that $u - v \in E_t(g)$ if, and only if $u \neq v$ and $u, v$ have at least one tag in common[7] in all their photos (including photos not in group $g$: the idea is to consider links between users independently of the groups).

## 2.3 Measuring proximity between tag clouds

In order to add a weight function to thematic edges (taking into account the thematic proximity of two users), we need some definitions on tags:

- $n_t$ (resp $n_t(u)$), with $t \in T$ and $u \in U$, is the number of photos (resp. photos of user $u$) having tag $t$, including photos outside studied groups;
- $n_{\max}$ is $\max_{t' \in T} n_{t'}$, the maximal number of photos having a tag;
- the *rarity coefficient* of a tag $t$ is defined by: $\rho_t = log(1 + \frac{n_{\max}}{n_t})$. This coefficient ranges from 1 for the most used tag *beach* to approximately 10 for the rarest ones;
- the *tag-weight* of tag $t$ on user $u$ is defined by: $w_{u,t} = \begin{cases} 0 & \text{if } n_t(u) = 0 \\ 1 + \log\, n_t(u) & \text{otherwise.} \end{cases}$
  The idea of the *log* is of course to reduce the impact of users posting thousands of photos about the same topic (their wedding, baby, cat, holiday...);
- the *edge weight* between users $u$ and $v$ is defined by:
  $w_{u,v} = w_{v,u} = \sum_{t \in T} (\rho_t \times \min\{w_{u,t}, w_{v,t}\})$, which is meant to tell whether $u$ and $v$ share many tags, taking into account the rarity of these tags: the rarer are the tags, the closer the users are to each other.

---

[3] `www.flickr.com`

[4] If for some reason you don't like one existing group on a subject, you just create a new one and recommend it to your contacts.

[5] This empirical choice was made in order to get a sample of groups that would have enough activity (not too small a size) and comparable sizes (much bigger sizes would have been more heterogeneous.)

[6] Of course these criteria are used as a proxy of social relations. In many circumstances, the contact functionality is used as a bookmark to a user's photos. This information may thus also indicate thematic relation.

[7] Here again, this is a proxy. Some tags are created by a particular community of users (*cc100*, *deleteme1*, *top-f25*...) and could thus be seen as social indicators.

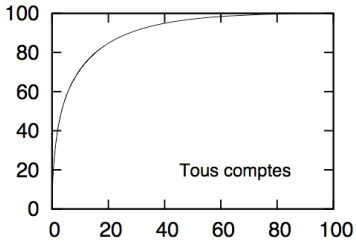
Of course pairs of users sharing at least one tag are numerous, which makes our thematic graphs much denser than social ones even though still sparse. Computing the edge weights is thus the heavier step of the whole analysis process.

# 3 Analysis

We will now define two indicators, one tag-oriented, one social-oriented, and give some ideas of more precise descriptions of groups.

## 3.1 Gini coefficient

Before describing the indicators, let us recall that a *Lorentz curve* graphically shows a cumulative distribution function. As an example, the figure shows the cumulative distribution of photos in the whole database, where the first 10% (resp. 60'%) of the users own 70% (resp. 98%) of the photos.
The *Gini coefficient* of a distribution is the area between the Lorentz curve and the diagonal (which is the Lorentz curve of the uniform distribution).




**Lorentz curve of photos ($x$-axis) owned by users ($y$-axis)**

This coefficient is a measure of the heterogeneity of the distribution: on the example, the highest numbers of photos owned by individuals are very high in comparison to photos owned by average people, the curve is thus far from the diagonal, the Gini coefficient is thus high.

## 3.2 Social and thematic indicators

**The social density** of a group $g$ is the *density* of the social graph $G_s(g)$, *i.e.* the ratio of the number of actual edges by the number of possible edges given the number of vertices. A relatively **high** social density indicates a **great amount** of 'social activity' between members of $g$.

**The interest-sharing heterogeneity** of a group $g$ is the Gini coefficient of the distribution of edge weights of $G_t(g)$. A relative **low value** indicates a homogeneous distribution, which means that average members of the group have as many tags in common, thus indicating **thematic concentration**.

Of course other indicators could be used to enrich the latter, *e.g.* for social activity, *clustering coefficient* of the social graph, or for thematic concentration, the Gini coefficient of a distribution of the representativity of the group for all tags[8]. The actual Lorentz curves could also be used, different curve shapes indicating different types of groups.

# 4 Results

Next page's chart shows the diversity of values for all groups on both social and thematic indicators.

A first interesting thing is to look at the most thematic groups, whose position is in the lower part of the chart. They are listed on the left-hand side of the chart. Three-quarters of these group are in two categories: geographical, especially cities (Buenos Aires, Tel Aviv, Taipei etc.) and technical groups (K750i, XPRO, Fuji etc.), whose social densities range from very low values (Vienna, Stockholm for cities, K750i, expired films for technical) to quite high ones (Tel Aviv, Buenos Aires and toycamera, XPRO).

As for groups with high social density, listed on the right-hand side of the chart, let us discuss on the first three that appear on the far right on the chart. The group *Paralelas/Parallels* is intended for photos with parallel lines (wires, skyscrapers etc.), which could mean any kind of photos (the interest-sharing heterogeneity is high). But as suggested by the title in Portuguese, many members are from Brazil. This is an example of a social group whose social activity comes from a geographical proximity of its members[9]. The group *FLICKRGAYS* is one of the (quite few) examples of both thematic *and* social groups[10] and is of

[8] The representativity of a group $g$ for a tag $t$ could be defined by the ratio of the sums of tag-weights of $t$ on all users of $g$ among all *Flickr* users.

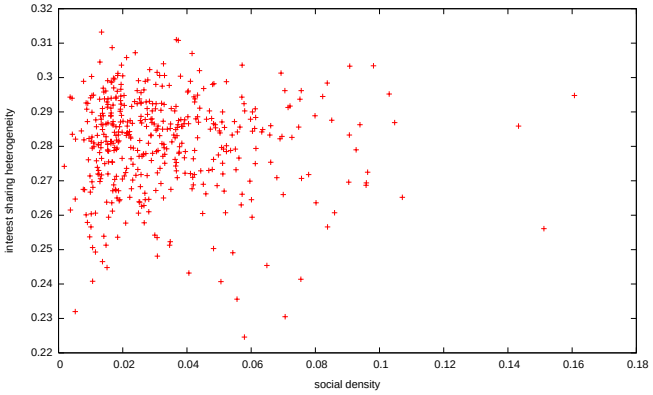
[9] This is not necessarily the case for the city groups mentionned above, that may contain many touristic pictures.

[10] in our two lists, these groups are FLICKRGAYS and toycamera.com.

course a real community. *Fifty Faves* is for photos having been marked as favorites by at least fifty users. Of course not thematic, this group is for very experienced *Flickr* users, who know each other and have discussions about their productions. In short, there is a wide range of these 'social' groups, whose names and declared purposes don't necessarily tell they are social.

**Most thematic groups**

| soc | thm | Group name |
|---|---|---|
| 0.06 | 0.22 | Buenos Aires |
| 0.07 | 0.23 | Tel Aviv Stories |
| 0.01 | 0.23 | K750i |
| 0.06 | 0.24 | Rio de Janeiro |
| 0.05 | 0.24 | Taipei |
| 0.01 | 0.24 | ROCKCLIMBING |
| 0.08 | 0.24 | toycamera.com |
| 0.04 | 0.24 | Hasselblad |
| 0.02 | 0.24 | Dublin |
| 0.06 | 0.25 | Philippines Images |
| 0.01 | 0.25 | Stockholm |
| 0.03 | 0.25 | Lisboa |
| 0.05 | 0.25 | XPRO CROSS PR |
| 0.01 | 0.25 | Vienna |
| 0.05 | 0.25 | Noir & Blanc |
| 0.01 | 0.25 | Pugs |
| 0.03 | 0.25 | I Shoot Fuji |
| 0.01 | 0.25 | Incredible India |
| 0.03 | 0.25 | expired film |
| 0.03 | 0.25 | Hamburg |
| 0.02 | 0.25 | Manchester. UK |
| 0.01 | 0.25 | snowboard |
| 0.01 | 0.25 | Panoramic |
| 0.03 | 0.25 | Baltimore |
| 0.15 | 0.26 | FLICKRGAYS |




**Social and thematic indicators for 450 groups**
**soc** is for social density, **thm** for interest sharing heterogeneity

**Most social groups**

| soc | thm | Group name |
|---|---|---|
| 0.15 | 0.26 | FLICKRGAYS |
| 0.14 | 0.29 | Fifty Faves |
| 0.11 | 0.27 | Pasted Paper |
| 0.10 | 0.29 | Nice Package! |
| 0.10 | 0.30 | Ethnic |
| 0.10 | 0.30 | poles / postes |
| 0.10 | 0.27 | Only 1000 |
| 0.10 | 0.27 | handbags |
| 0.10 | 0.27 | mens fashion & |
| 0.09 | 0.29 | Let thr be light! |
| 0.09 | 0.28 | Desaturado |
| 0.09 | 0.30 | Super Colored |
| 0.09 | 0.28 | blue/black/white |
| 0.09 | 0.27 | 6000+ Views |
| 0.09 | 0.26 | Zona libre |
| 0.09 | 0.29 | Male Butt Crack |
| 0.08 | 0.26 | Iconia Fashion |
| 0.08 | 0.30 | "*Alba luminis |
| 0.08 | 0.29 | 4Variations |
| 0.08 | 0.26 | Handmade |
| 0.08 | 0.29 | Painterly |
| 0.08 | 0.27 | Toy Face |
| 0.08 | 0.27 | Photoheart |
| 0.08 | 0.30 | Verde - Green - Vert |
| 0.08 | 0.24 | toycamera.com |

Besides showing the great diversity of uses of *Flickr* groups, these empirical results suggest that the methodological scheme presented in this paper may indeed be used in order to detect groups having a presumably strong social and/or thematic 'identity'. This could serve many purposes like targetting specific communities for designing of new services, studying how to make thematic groups become social etc.

Future work should address the issue of comparing social and thematic indicators for groups of various sizes as well as studying ways of dealing efficiently with the whole database (70,000 groups with up to 30,000 members), for instance by finding heuristics to avoid computing all edge weights.

# References


[AHA07] N. Ali-Hasan and L. Adamic. Expressing social relationships on the blog through links and comments. In *Intern. Conf. on Weblogs and Social Media*, 2007.

[BHK98] J. S. Breese, D. Heckerman, and C. Kadie. Empirical analysis of predictive algorithms for collaborative filtering. In *Uncertainty in AI. Proc. of 14th conf.* Morgan Kaufman, 1998.

[BP98] Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, 30(1–7):107–117, 1998.

[CJPCP06] D. Cardon, V. Jeanne-Perrier, F. Le Cam, and N. Pelissier, editors. *Autopublications*, volume 24. Réseaux, 2006.

[FLG00] Gary Flake, Steve Lawrence, and C. Lee Giles. Efficient identification of web communities. In *Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 150–160, Boston, MA, 2000.

[LP04] Matthieu Latapy and Pascal Pons. Computing communities in large networks using random walks. Technical report, arXiv.org, 2004.

[LWGS06] T. Lento, H. Welser, L. Gu, and M. Smith. The ties that blog: Examining the relationship between social ties and continued participation in the wallop weblogging system. In *3rd annual workshop on the Weblogging Ecosystem*, Edimburgh, 2006.

[Mil67] S. Milgram. The small world problem. *Psychology Today*, pages 60–67, 1967.

[New04] M. Newman. Fast algorithm for detecting community structure in networks. *Physical Review E*, 69, 2004.

[Wat03] Duncan Watts. *Six Degrees: The Science Of A Connected Age*. W.W.Norton, London, 2003.

[WBB76] H. White, S. Boorman, and R. Breiger. Social structure for multiple networks i. blockmodels of roles and positions. *American Journal of Sociology*, 81, 1976.

[WF94] S. Wasserman and K. Faust. *Social Network Analysis: Methods and Applications.* Cambridge University Press, 1994.